

Occupational Requirements Survey: results from a job observation pilot test

The Bureau of Labor Statistics (BLS) conducts the Occupational Requirements Survey (ORS), an establishment survey, for the Social Security Administration. The survey collects information on the vocational preparation for and the cognitive and physical requirements of occupations in the U.S. economy, as well as the environmental conditions in which those jobs are performed. BLS field economists, who conduct interviews with establishment representatives, collect these data. On the basis of stakeholders asking whether the data collected through this mode would result in similar measurements as data collected through direct job observation, BLS conducted a job observation pilot test during the summer of 2015. As part of this test, field economists recontacted establishments who had responded to the ORS preproduction test, and while observing workers performing their jobs, the field economists obtained data on the physical job requirements. The results showed relatively high rates of agreement between observed and collected data for most physical requirements.

In the summer of 2012, the Social Security Administration (SSA) and the Bureau of Labor Statistics (BLS) signed an interagency agreement to begin testing the collection of occupational requirements data. As a result, BLS established the Occupational Requirements Survey (ORS) as a test survey in late 2012. The goal of ORS is to collect and publish occupational information that meets the needs of SSA at the level of an eight-digit code based on the Standard Occupational Classification (SOC) system that is used by the Occupational Information Network (O*NET).¹



Tiffany Y. Chang

chang.tiffany@bls.gov

Tiffany Y. Chang is a statistician in the Office of Employment and Unemployment Statistics, U.S. Bureau of Labor Statistics.

Kristen A. Monaco

Monaco.Kristen@bls.gov

Kristen A. Monaco is the Associate Commissioner of the Office of Compensation and Working Conditions, U.S. Bureau of Labor Statistics.

Kristin Smyth

smyth.kristin@bls.gov

Kristin Smyth is a labor economist in the Office of Compensation and Working Conditions, U.S. Bureau of Labor Statistics.

The ORS data are collected under the umbrella of the National Compensation Survey, which field economists use to collect data. Field economists generally collect data elements either through a personal visit to the establishment or remotely by telephone, email, mail, or a combination of modes.

For ORS, field economists collect occupationally specific data elements to meet the needs of the SSA in the following categories:

- Physical demands
- Specific vocational preparation
- Mental and cognitive demands
- Environmental conditions in which the work is performed

In fiscal year 2015, BLS completed data collection for the ORS preproduction test. The preproduction test might better be described as a “dress rehearsal” because the collection procedures, data-capture systems, and review were structured to be as close as possible to the structure of those that would be used in production.²

Background on job observation pilot test

The ORS job observation pilot test was intended to assess whether the data collected through ORS interview collection methods are systematically different from the data collected through direct observation. This test was conducted in response to both public comments in a *Federal Register* notice and an external subject matter expert’s recommendations for testing and validating ORS design.³

The observation test was conducted in the summer of 2015, running from June to September. The observation test involved recontacting a subset of establishments that were interviewed as part of the preproduction test. Pairs of field economists were sent to observe select jobs within each establishment and record data on the physical and environmental data elements during a 1-hour observation period.⁴ The 1-hour observation period sought to achieve a balance between gathering data that represent the job and limiting the respondent burden involved in conducting such a test.

Because the goal of ORS is to produce estimates at the eight-digit O*NET SOC level, the observation test was structured to allow us to compare preproduction data with observed data at the eight-digit SOC level as well. Thus, a subset of occupations was chosen for inclusion in the test. The subset was chosen on the basis of two criteria:

1. The occupation must appear in the preproduction microdata at least 40 times.
2. The jobs have substantial physical requirements and are relatively simple to observe (e.g., long-haul truck drivers were excluded).

These criteria resulted in the following occupations sampled for the observation test:

- Nursing assistants
- Cooks, institution and cafeteria
- Cooks, restaurant
- Waiters and waitresses

- Dishwashers
- Janitors and cleaners
- Maids and housekeeping cleaners
- Childcare workers
- Cashiers
- Retail salespersons
- Receptionists and information clerks
- Team assemblers
- Industrial truck and tractor operators
- Laborers and freight, stock, and material movers, hand

Procedures for job observation pilot test

The sample consisted of 540 jobs (456 from private industry and 84 from state and local governments) from existing ORS preproduction establishments. Establishments were sampled across geography, industry, and size to ensure a good distribution of available establishments within each occupation.

For each of the sampled establishments and occupations, a field economist secured the appointment and explained to the respondent the reason for the followup visit. A pair of field economists then collected data by way of a personal visit. The paired field economists, if possible, simultaneously observed the same employee in a preselected job and documented the situation for a 60-minute period. However, if pairing was not possible, one economist would observe a separate employee. The field economists were instructed not to look at data recorded from the preproduction test for their establishments or to discuss their data with one another. Each field economist independently recorded and coded his or her observations during the personal visit. Field economists were as inconspicuous as possible and did not ask questions of the observed employee.

The field economists were instructed to code the duration in minutes and to code a duration of zero if the element was not observed. Some elements had additional questions, such as, Was the job performed with one hand or both? For these elements, each field economist checked the appropriate box category and noted the duration in minutes. Field economists then checked their data for accurate recording before marking the schedule and quote as complete. The team overseeing the test held two collection debrief meetings with the field economists (one midtest and one at the end of collection) to assess how the process worked.

Response rates

Field economists contacted establishments for 405 of 540 jobs in the sample and observed 244 jobs, a 60-percent response rate.⁵ As shown in table 1, the refusal rate varied by occupation.

Table 1. Response rates for Occupational Requirements Survey job observation test, by occupation

Occupation	Observed	Response rate
All occupations	244	60
Nursing assistants	9	31
Cooks, institution and cafeteria	19	61

See footnotes at end of table.

Table 1. Response rates for Occupational Requirements Survey job observation test, by occupation

Occupation	Observed	Response rate
Cooks, restaurant	16	59
Waiters and waitresses	19	66
Dishwashers	13	52
Janitors and cleaners	25	74
Maids and housekeeping cleaners	20	71
Childcare workers	6	37
Cashiers	22	67
Retail salespersons	17	57
Receptionists and information clerks	23	68
Team assemblers	17	71
Industrial truck and tractor operators	17	53
Laborers and freight, stock, and material movers, hand	21	64

Source: U.S. Bureau of Labor Statistics.

Childcare workers and nursing assistants had very high refusal rates. These refusals largely stemmed from establishments' concerns about privacy under state and national laws. Some successful observations of these two occupations occurred during the observation test; however, because of the small sample size, we did not include any childcare workers or nursing assistants in our test analysis.

Measures of agreement between preproduction and observed duration

We evaluated the agreement between the observed values of the data elements and those collected (during the interviews) in the preproduction test. We will refer to these as “observed” and “interview” values hereon.⁶ Our analysis focused on the physical elements defined in table 2.

Table 2. Description of physical elements in Occupational Requirements Survey

Physical demand	Description
Crawling	Moving about on hands and knees or hands and feet.
Crouching	Bending the body downward and forward by bending legs and spine.
Kneeling	Bending legs at knees to come to rest on knee(s).
Stooping	Bending the body downward and forward by bending the spine at the waist.
Reaching overhead	Extending hands and arms in a 150- to 180-degree vertical arc.
Reaching at or below shoulder level	Extending hands and arms from 0 up to 150 degrees in a vertical arc.
Communicating verbally	Expressing or exchanging ideas by means of the spoken word to impart oral information.
Keyboarding	Entering text or data into a computer or other machine by using a keyboard or other device. This element is measured separately for standard keyboards and for touchscreen, 10-key, and other keyboards.
Fine manipulation	Picking, pinching, or otherwise working primarily with fingers rather than the whole hand or arm.
Gross manipulation	Seizing, holding, grasping, turning, or otherwise working with hand(s).
Pushing or pulling	Exerting force upon an object so that it moves away from (pushing) or toward (pulling) the force. This element is measured separately for hands and arms, feet and legs, and feet.

See footnotes at end of table.

Table 2. Description of physical elements in Occupational Requirements Survey

Physical demand	Description
Climbing Ramps or stairs	Ascending or descending ramps and/or stairs by using feet and legs.
Climbing ladders, ropes, or scaffolding	Ascending or descending ladders, scaffolding, ropes, or poles and the like.
Source: U.S. Bureau of Labor Statistics.	

The durations of most physical elements for preproduction were classified into five categories:

1. Not present
2. Seldom—up to 2 percent of the day
3. Occasionally—2 percent up to one-third of the day
4. Frequently—one-third up to two-thirds of the day
5. Constantly—two-thirds or more of the day

Measuring agreement between observed and interview data was complicated by two factors:

1. The duration of the observation test was short, which may lead to discrepancies between the presence or absence of certain physical requirements. In particular, we expected high degrees of agreement in the presence or absence of physical requirements with durations that fall into the “frequently” or “constantly” categories and low degrees of agreement for elements that occur “occasionally” or “almost never.”
2. In preproduction collection, roughly 20 percent of the physical requirements that were classified as “present” in the job had no duration provided by the respondent. The unknown duration is especially high in particular elements—in the sample of jobs that were observed, the interview data had missing duration in nearly 30 percent of the cases for communicating verbally and 25 percent of the cases for fine manipulation.

To address the challenges posed by the short duration of the job observation test, we recategorized the durations into four categories, aggregating the classifications “not present” and “seldom” into one category:

1. Not present or seldom—less than 2 percent of the day
2. Occasionally—2 percent up to one-third of the day
3. Frequently—one-third up to two-thirds of the day
4. Constantly—two-thirds or more of the day

First, we calculated “raw” levels of agreement between the observed data and interview data. These levels are presented in column 2 of table 3. The levels of agreement are relatively high, ranging from 71.2 percent for reaching at or below shoulder level to 97.4 percent for pushing and pulling with feet.

Table 3. Percent agreement and Cohen's kappa measure of agreement between observed and interview data for duration of physical elements

Occupational Requirements Survey element	Agreement	Expected agreement	Cohen's kappa	p-value	PABAK
Crawling	97.1	97.1	−0.01	0.58	0.96
Crouching	79.3	77.8	.07	.18	.63
Kneeling	87.7	85.4	.16	<.01	.78
Stooping	74.0	71.7	.08	.02	.38
Reaching					
Overhead	84.3	81.0	.18	<.01	.62
At or below shoulder level	71.2	67.3	.12	<.01	.31
Communicating verbally	75.6	67.3	.25	<.01	.41
Keyboarding	92.1	81.5	.58	<.01	.81
Touchscreen	93.9	88.4	.47	<.01	.85
10-key	96.5	94.9	.32	<.01	.94
Other	95.9	94.8	.20	<.01	.90
Manipulation					
Fine	76.7	71.4	.19	<.01	.44
Gross	76.3	70.5	.21	<.01	.44
Pushing or pulling					
With hands and arms	73.6	66.6	.21	<.01	.37
With feet and legs	79.8	76.0	.16	<.01	.52
With feet	97.4	97.4	−.01	.57	.94
Climbing					
Ramps or stairs	89.8	88.6	.11	.05	.82
Ladders, ropes, or scaffolding	95.8	94.9	.17	<.01	.92

Note: PABAK = prevalence-adjusted and bias-adjusted kappa.
Source: U.S. Bureau of Labor Statistics.

The levels of agreement reported in table 3 suggest that the durations of physical requirements were similar across modes of collection, but statistical tests of agreement were required to ensure that the levels of agreement did not occur by chance. Therefore, we used a weighted version of Cohen's kappa to assess agreement across modes of collection.⁷ The kappa statistic measures the agreement against a benchmark of the expected agreement. Bear in mind, however, that if only a few possible categories exist, the observed and interview data could fall into the same duration category simply by chance. The weighting in our kappa measure penalized less for disagreements that were close (e.g., the observed duration was “frequently” and the duration from the interview was “constant”) and more severely for disagreements that were farther apart (e.g., the observed duration was “not present or seldom” and the duration from the interview was “constantly”).

Kappa generally ranges from −1 to +1. Negative values of kappa indicate that the level of agreement is less than the expected agreement. Kappa statistics close to 1 imply a higher level of agreement. Although some controversy exists in the literature regarding thresholds of kappa, J. Richard Landis and Gary Koch have proposed the following standards:⁸

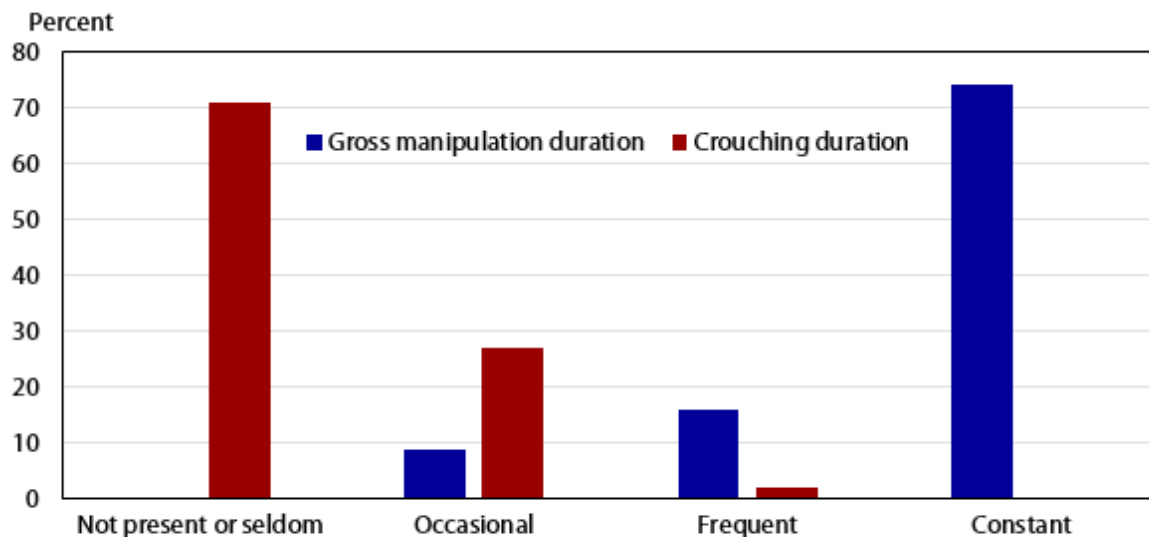
- Less than or equal to 0 is poor agreement.

- 0.01–0.20 is slight agreement.
- 0.21–0.40 is fair agreement.
- 0.41–0.60 is moderate agreement.
- 0.61–0.80 is substantial agreement.
- 0.81–1.00 is almost perfect agreement.

As can be seen in column 3 of table 3, however, the expected levels of agreement are relatively high and the kappa statistics are relatively low.⁹ With the exception of crawling, crouching, and pushing and/or pulling with feet, the agreement levels are greater than the expected levels of agreement in a 5-percent one-tailed test (column 5 of table 3). The kappa statistics vary considerably; the average value of the kappa statistic is 0.20, which denotes relatively low levels of agreement relative to agreement by chance.

A well-known issue with kappa is the influence of prevalence and bias on the kappa measures. Generally, variables with underlying uniform distributions result in higher values of kappa. Measuring kappa using data that have skewed prevalence can cause the “kappa paradox,” in which high levels of rater agreement have relatively low kappa statistics.¹⁰ The distributions of the physical elements in ORS, however, tend to be highly skewed (in terms of both underlying discrete values of the duration as well as the categorical measures). As figure 1 illustrates, the distribution of observed duration for two physical elements—crouching and gross manipulation. The mode for crouching is “not present or seldom,” which was recorded in over 70 percent of the observed jobs. Frequent or constant crouching rarely occurs. On the other end of the spectrum, gross manipulation is constant for most of the observed jobs, which is not surprising when one considers the occupations sampled (cashier, maid, etc.).

Figure 1. Job observation distribution of gross manipulation and crouching durations



Note: A measure of prevalence-adjusted and bias-adjusted kappa was used to account for the skewness in data and is presented in table 3, column 6.

Source: U.S. Bureau of Labor Statistics.

To reduce the impact of skewness on our test statistics, we used a measure of prevalence-adjusted and bias-adjusted kappa (PABAK). The PABAK measure is presented in column 6 of table 3. The average is 0.68, which

is considered “substantial”; however, the PABAK measure varies considerably by data element. In particular, stooping, reaching at or below the shoulder, communicating verbally, fine manipulation, gross manipulation, and pushing and/or pulling with hands and arms have low measures (less than 0.50), even after prevalence and bias are adjusted.

Of particular concern, given the potential uses of ORS data in the disability determination process, was whether the preproduction interview data appear to understate the duration of the physical elements, especially for those elements with relatively low PABAK measures. We evaluated potential under- or overstatement of duration using a sign test. The sign test is a test of the difference in medians. We were particularly interested in elements in which the sign test rejects the null hypothesis that the observed median is less than or equal to the interview median—rejecting this hypothesis implies that the observed durations were the durations collected through interviews (see column 2 of table 4).

Table 4. P-values associated with test of differences in median values between interview and observation

Occupational Requirements Survey element	Observed median of null hypothesis	
	Less than interview	Greater than interview
Crawling	0.66	0.66
Crouching	.50	.59
Kneeling	.97	.05
Stooping	<.01	1.00
Reaching		
Overhead	.81	.25
At or below shoulder level	<.01	1.00
Communicating verbally	.95	.07
Keyboarding	.68	.44
Touchscreen	.35	.79
10-key	.29	.87
Other	<.01	1.00
Manipulation		
Fine	<.01	1.00
Gross	<.01	1.00
Pushing and/or pulling		
With hands and arms	<.01	1.00
With feet and legs	.22	.84
With feet	.02	1.00
Climbing		
Ramps or stairs	.98	.05
Ladders, ropes, or scaffolding	.96	.21

Source: U.S. Bureau of Labor Statistics.

The elements with longer duration associated with observation are stooping, reaching at or below the shoulder, other keyboarding, fine manipulation, gross manipulation, pushing and/or pulling with hands and arms, and pushing and/or pulling with feet. When we measured the modes of these elements, only one differed in mode

between the collected and observed values—reaching at or below the shoulder level. The value of the mode for this element was “occasionally” (2 percent up to one-third of the day) in the interview data and “constantly” in the job observation data (two-thirds or more of the day).

Missing duration was identified as an issue with ORS preproduction data for some of the physical elements. In the case of reaching at or below the shoulder level, 53 of the job observation duration measures could not be compared with interview duration data because of missing duration. Notably, among the 53 missing quotes in preproduction, 64 percent of the quotes of the job observation test recorded durations of frequently or constantly. This result was a common pattern among those elements in which the sign test rejected the null of observation duration equal to or below preproduction—the missing data in preproduction aligned with observed durations above the median and mode.

Conclusions

The job observation pilot test validated the ORS physical elements by comparing the data collected through preproduction interviews with those data collected through a different source—observation. Pairs of field economists were assigned to observe the same job for 60 minutes and record the duration of each physical element of the job.

For elements that workers performed infrequently, the 60-minute observation period may have led to more disagreement between observed data and interview data collected during preproduction. The PABAK measures of duration were relatively strong, suggesting that the data collected during interviews and observations had high levels of agreement across most elements.

NOTES

¹ Most typically used by BLS, the Standard Occupational Classification (SOC) system uses six-digit codes (<https://www.bls.gov/soc/>), generally referred to as “detailed occupations.” O*NET uses a more detailed occupational taxonomy (<https://www.onetcenter.org/taxonomy.html>), classifying occupations by eight-digit codes and referring to these as “O*NET-SOC 2010 occupations.” The SOC has 840 six-digit and 1,110 eight-digit codes.

² The sample design was similar to what will be used in production but was altered to meet test goals. A report on the Occupational Requirements Survey preproduction test is available at <https://www.bls.gov/ncs/ors/pre-prod-report.pdf>.

³ A link to the subject matter expert’s report can be found at https://www.bls.gov/ncs/ors/pre-prod_estval.pdf.

⁴ The field economists used devices to collect data on a subset of the Occupational Requirements Survey environmental elements. Unfortunately, problems with the readings of some devices resulted in the analysis of those elements being dropped from this research.

⁵ The remaining 135 jobs were at establishments that were not contacted during the time of the test.

⁶ We also assessed the agreement between the paired field economists. The analysis of interrater agreement is available in the “Occupational Requirements Survey job observation report” at https://www.bls.gov/ncs/ors/preprod_job_ob.pdf.

[7](#) John Cohen, “Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit,” *Psychological Bulletin*, vol. 70, no. 4, 1968, pp. 213–220.

[8](#) J. Richard Landis and Gary G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, no. 1, March 1977, pp. 159–174, https://www.jstor.org/stable/2529310?seq=1#page_scan_tab_contents.

[9](#) Measures of expected agreement, kappa statistics, and the standard errors needed to compute p -values were calculated with use of Stata (version 12). Stata’s calculations are based on J. Richard Landis and Gary G. Koch, “A one-way components of variance model for categorical data,” *Biometrics*, vol. 33, no. 4, December 1977, pp. 671–679, https://www.jstor.org/stable/2529465?seq=1#page_scan_tab_contents; and standard errors are based on Joseph L. Fleiss, John C. Nee, and J. Richard Landis, “Large sample variance of kappa in the case of different sets of raters,” *Psychological Bulletin*, vol. 86, no. 5, September 1979, pp. 974–977, <http://psycnet.apa.org/psycinfo/1979-32706-001>.

[10](#) For more information, see Domenic V. Cicchetti and Alvan R. Feinstein, “High agreement but low kappa II: resolving the paradoxes,” *Journal of Clinical Epidemiology*, vol. 43, no. 6, 1990, pp. 551–558, <https://www.ncbi.nlm.nih.gov/pubmed/2189948>; and Alvan R. Feinstein and Domenic V. Cicchetti, “High agreement but low kappa I: the problem of two paradoxes,” *Journal of Clinical Epidemiology*, vol. 43, no. 6, 1990, pp. 543–549, <https://www.ncbi.nlm.nih.gov/pubmed/2348207>.

RELATED CONTENT

Related Articles

[Which industries need workers? Exploring differences in labor market activity](#), *Monthly Labor Review*, January 2016.

[Industry employment and output projections to 2024](#), *Monthly Labor Review*, December 2015.

[The Occupational Requirements Survey: estimates from preproduction testing](#), *Monthly Labor Review*, November 2015.

Related Subjects

[Sampling](#) | [Time use](#) | [Survey methods](#) | [Productivity](#)